Advancing Metrics on the Standards Track:

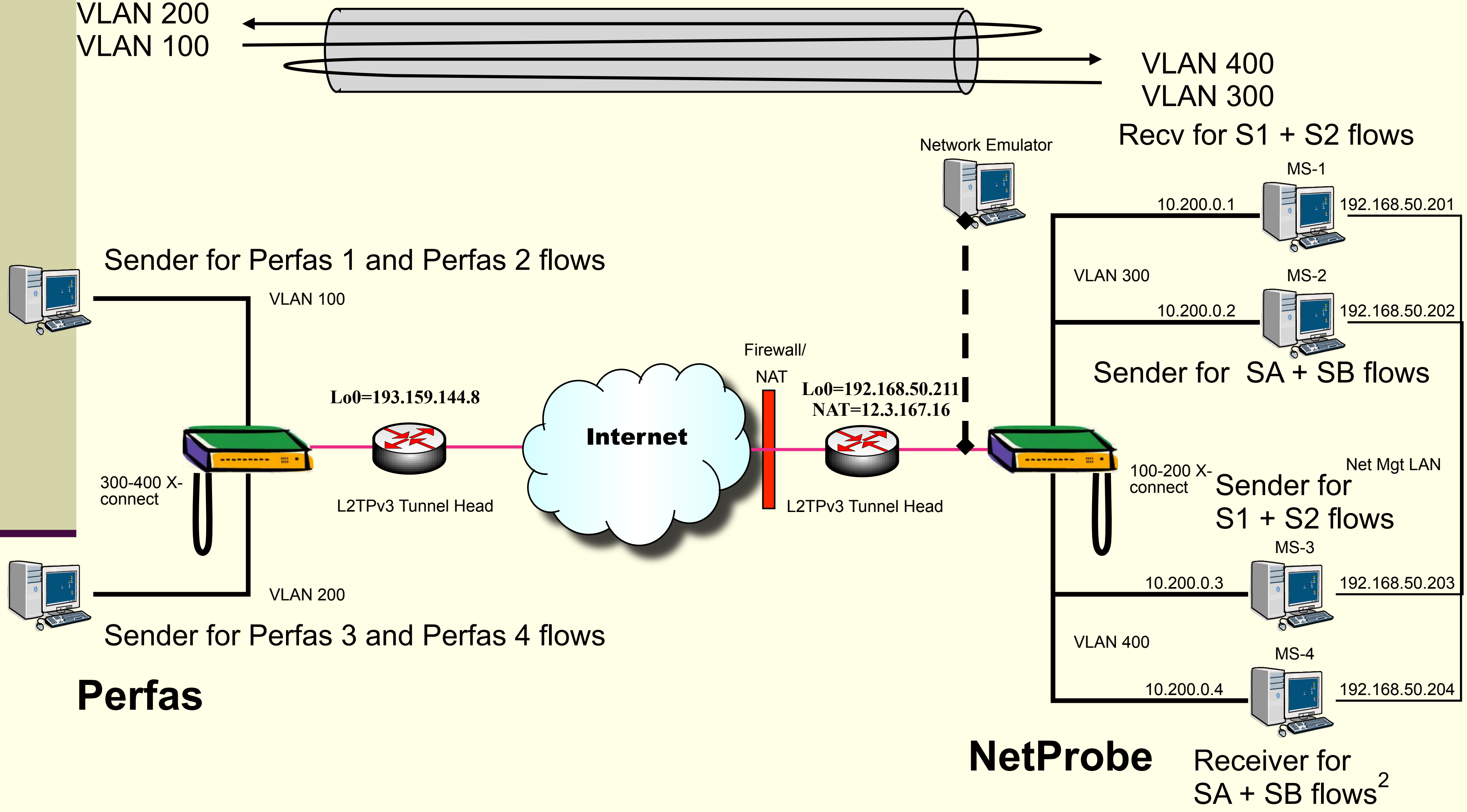
RFC 2679 Test Plan and Results

draft-morton-ippm-testplan-rfc2679-01

Len Ciavattone, Rüdiger Geib,
Al Morton, Matthias Wieser

March 2012

Test Configuration



VLAN 200
VLAN 100

VLAN 400
VLAN 300

Recv for S1 + S2 flows

Network Emulator

MS-1

10.200.0.1

192.168.50.201

Sender for Perfas 1 and Perfas 2 flows

VLAN 100

VLAN 300

MS-2

10.200.0.2

192.168.50.202

Firewall/
NAT

Sender for SA + SB flows

Lo0=193.159.144.8

Lo0=192.168.50.211
NAT=12.3.167.16

Internet

300-400 X-
connect

L2TPv3 Tunnel Head

L2TPv3 Tunnel Head

100-200 X-
connect

Net Mgt LAN

Sender for
S1 + S2 flows

MS-3

10.200.0.3

192.168.50.203

VLAN 200

Sender for Perfas 3 and Perfas 4 flows

VLAN 400

MS-4

10.200.0.4

192.168.50.204

Perfas

NetProbe

Receiver for
SA + SB flows

Tests in the Plan

- 6. Tests to evaluate RFC 2679 Specifications
  - **6.1. One-way Delay, ADK Sample Comparison – Same & Cross Implementations** <<< Additional test results
  - 6.2. One-way Delay, Loss threshold,
  - 6.3. One-way Delay, First-bit to Last bit,
  - 6.4. One-way Delay, Difference Sample Metric
  - 6.5. Implementation of Statistics for One-way Delay

Overview of Testing

- 32 different experiments conducted from March 9 through May 2, 2011.
- Varied Packet size, Active sampling distribution, test duration, and other parameters (Type-P)
- Added Network Emulator “netem” and varied fixed and variable delay distributions
  - This talk describes tests beyond 100ms+/-50
  - Also inserted loss in a limited number of experiments.

Overview of Additional Testing

- The common parameters used for tests in this section are:
- o IP header + payload = 64 octets
- o Periodic sampling at 1 packet per second
- o Test duration = 300 seconds at each delay variation setting for a total of 1200 seconds (May 2, 2011 at 1720 UTC)

- The netem emulator was set for 100ms average delay, with (emulated) uniform delay variation of:
- o +/-7.5 ms
- o +/-5.0 ms
- o +/-2.5 ms
- o 0 ms

Results for May 2 tests

Emulated Delay		Sub-Sample size		
Variation 0ms				
adk.combined (all)	300 values		75 values	
Adj. for ties	raw	mean adj	raw	mean adj
TC observed	226.6563	67.51559	54.01359	21.56513
P-value	0	0	0	0
Mean std dev (all),us	719		635	
Mean diff of means,us	649	0	606	0
Variation +/- 2.5ms				
adk.combined (all)	300 values		75 values	
Adj. for ties	raw	mean adj	raw	mean adj
TC observed	14.50436	-1.60196	3.15935	-1.72104
P-value	0	0.873	0.00799	0.89038
Mean std dev (all),us	1655		1702	
Mean diff of means,us	471	0	513	0

Results for May 2 tests (contd.)

Emulated Delay	Sub-Sample size			
Variation +/- 5ms				
adk.combined (all)	300 values		75 values	
Adj. for ties	raw	mean adj	raw	mean adj
TC observed	8.29921	-1.28927	0.37878	-1.81881
P-value	0	0.81601	0.29984	0.90305
Mean std dev (all),us	3023		2991	
Mean diff of means,us	582	0	513	0
Variation +/- 7.5ms				
adk.combined (all)	300 values		75 values	
Adj. for ties	raw	mean adj	raw	mean adj
TC observed	2.53759	-0.72985	0.29241	-1.15840
P-value	0.01950	0.66942	0.32585	0.78686
Mean std dev (all),us	4449		4506	
Mean diff of means,us	426	0	856	0

Results

- 1. None of the raw or mean adjusted results pass the ADK criterion with 0 ms emulated delay variation. Use of the 75 value sub-sample yielded the same conclusion. (We note the same results when comparing same implementation samples for both NetProbe and Perfas.)

- 2. When the smallest emulated delay variation was inserted (+/-2.5ms), the mean adjusted samples pass the ADK criterion and the high P-value supports the result. The raw results do not pass.

- 3. At higher values of emulated delay variation (+/-5.0ms and +/-7.5ms), again the mean adjusted values pass ADK. We also see that the 75-value sub-sample passed the ADK in both raw and mean adjusted cases. This indicates that sample size may have played a role in our results, as noted in the Appendix of [RFC2680] for Goodness-of-Fit testing.

BACKUP

Backup Backup Backup

Section 6.1 One-way Delay, ADK Sample Comparisons (Same/Cross)

1. Configure tests on an L2TPv3 tunnel over a live network path.
2. Measure a sample of one-way delay singletons with 2 or more implementations, using identical options.
3. Measure a sample of one-way delay singletons with *four* instances of the *same* implementations,
   - connectivity differences SHOULD be the same as for the *cross* implementation tests.
4. Apply ADK comparison: same (see App C of metrictest)
5. Take coarsest confidence/resolution, or Section 5 Limits
6. Apply constant correction factors (Section 5)
7. Compare Cross-Implementation ADK for equivalence (samples come from same distribution)

Criteria for the Equivalence Threshold and Correction Factors

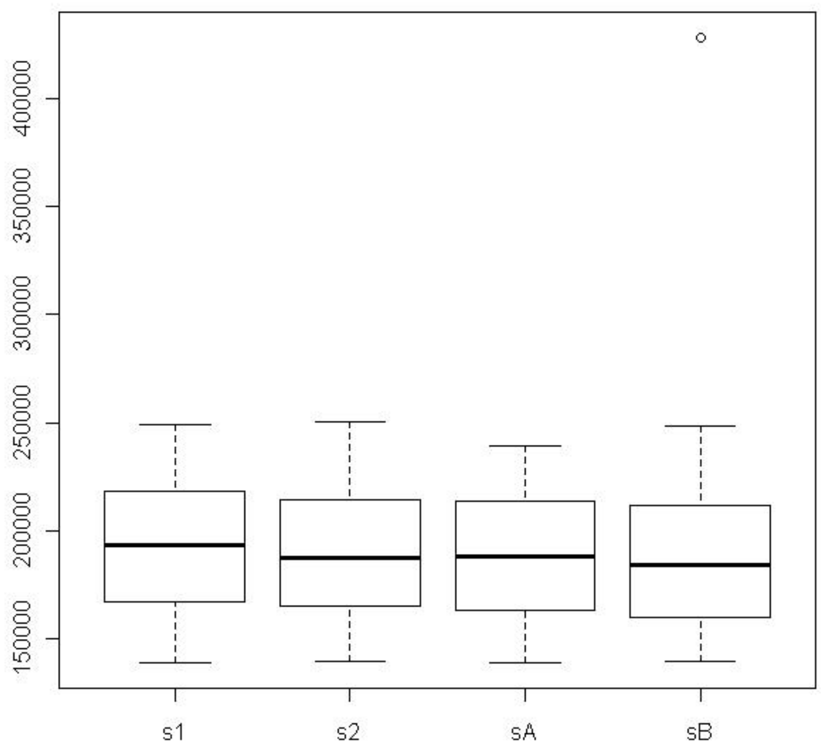
- Purpose: Evaluate Specification Clarity (using results implementations)
- For ADK comparison: cross-implementations
  - 0.95 confidence factor at 1ms resolution, or
  - The smallest confidence factor & res. of *same* Imp.
- A constant time accuracy error < +/-0.5ms MAY be removed from one Implementation before ADK or comparison of means
- A constant propagation delay error < +2ms MAY be removed from one Implementation …
  - (due to use of different sub-nets between the switch and measurement devices at each location)

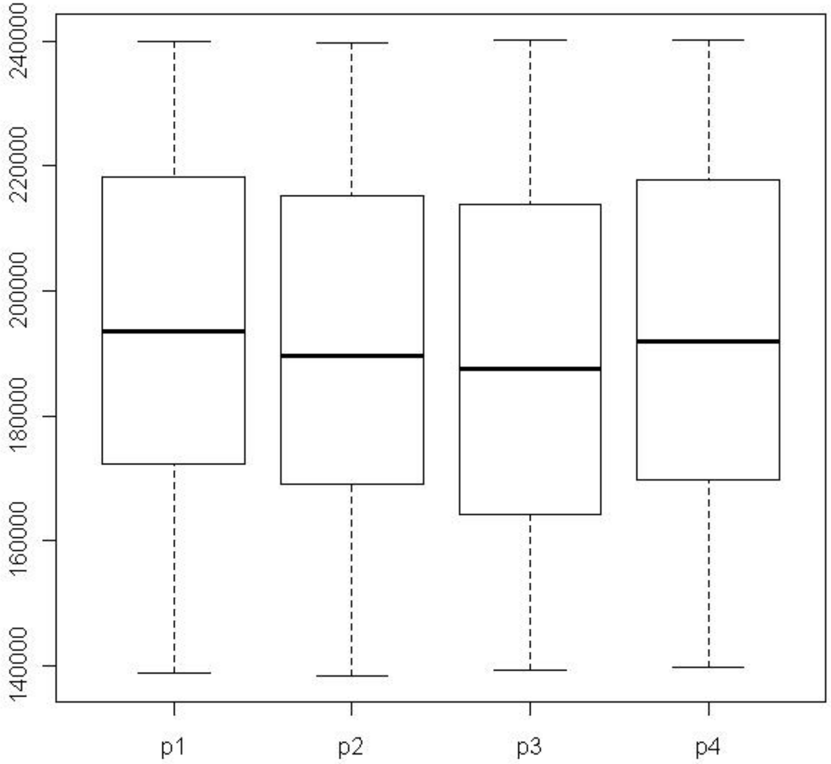
Overview of Testing (sample)

Date	Samp	Interval	Duration	Notes	ADK same	ADK cross
Mar 23	Poisson	1s	300s	Netem 10% Loss		
Mar 24	Periodic	1s	300s	Netem 100ms +/- 50ms delay		
Mar 24	Periodic	1s	300s	Netem 10% Loss		
Mar 28	Periodic	1s	300s	Netem 100ms		
Mar 29	Periodic (rand st.)	1s	300s	Netem 100ms +/- 50ms delay, 64 Byte	NP s12AB Per p1234	Pass combined
Apr 6	Periodic (rand st.)	1s	300s	Netem 100ms +/- 50ms delay, 340 Byte		
Apr 7	Periodic (rand st.)	1s	1200s	Netem 10% Loss		
Apr 12	Periodic (rand st.)	1s	300s	Netem 100ms, 500 Byte and 64 Byte comparison		

Summary of March 29 Tests No correction factors used, 1usec res.

- NetProbe
- Perfas+





ADK tests – Glossary & Background

The ADK R-package returns some values and these require interpretation:

ti.obs is calculated, an observed value based on an ADK metric. The absolute ti.obs value must be less than or equal to the Critical Point.

The P-value or (P) in the following tables is a statistical test to bolster confidence in the result. It should be greater than or equal to α = 0,05.

Critical Points for a confidence interval of 95% (or α = 0.05)
For k = 2 samples, the Critical Point is 1.960
For k = 4 samples, the Critical Point is 1.915
For k = 9 samples, the Critical Point is 1.839
(Note, the ADK publication doesn't list a Critical Point for 8 samples, but it can be interpolated)

Green = ADK test passed, Red = ADK test failed

ADK for Mar 29 tests – Perfas+

```
| ti.obs (P) |  perfas 1   |  perfas 2   |  perfas 3   |
|            |             |             |             |
.............|.............|.............|.............|
|            |             |             |             |
|   perfas 2 |
|            |             |             |             |
|   perfas 3 |

|            |             | 0.37 (0.24) |             |
|...perfas.3.|.1.09.(0.12).|.............|.............|


                                         |
+Perfas-ADK-Results-for-same-0.13-(0.37)-+-1.36-(0.09)-+
```

Green = passed, Red = failed

Perfas ADK Results for same-implementation

Green = passed, Red = failed

ADK for Mar 29 – Cross-Implementations

```
Null Hypothesis:
  All samples within a data set come from a common distribution.
```

All samples within a data set come from a common distribution.

```
  All NetProbe combined          ti.obs         P-value
  Not adj. for ties              0.64999        0.21355
  adj. for ties
                                 0.64833        0.21392

  All Perfas combined            0.35968        0.23412
  Not adj. for ties
  adj. for ties

  All Netprobe and Perfas combined
  Not adj. for ties              0.85537        0.17967
  adj. for ties
```

Other Results (details in the memo)

- Calibration – completed for both implementations
- Loss Threshold – available in post-processing for both implementations
  - Suggest revised text to allow implementations
  - Loss Threshold – available in post-processing for both implementations
- First bit – Last bit – issues with test design
  - Low speed links not available
  - Emulator interfaces found in Half-Duplex
  - Replace with description of this in RFC
- Differential Delay – sufficiently accurate
- Delay Stats – drop Percentile in this RFC
- First bit – Last bit – issues with test design
  - Low speed links not available
  - Emulator interfaces found in Half-Duplex

Summary

Test Plan for Key clauses of RFC 2679

- the basis of Advance RFC Request
  - Criteria for Equivalence Threshold & correction factors

Adopt as a WG document?

- Experiments complete, key clauses of RFC2679 evaluated

Two revisions to the RFC suggested from this

- Two revisions to the RFC suggested from this study

References

R Development Core Team (2011), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Scholz F.W. and Stephens M.A. (1987), K-sample Anderson-Darling Tests, *Journal of the American Statistical Association,* **Vol 82, No. 399**, 918–924.

[Table 1 of Scholz and Stevens]

m (k-1)	0.75 α=0.25	0.90 α=0.1	0.95 α=0.05	0.975 α=0.025	0.99 α=0.01
1	.326	1.225	1.960	2.719	3.752
2	.449	1.309	1.945	2.576	3.414
3	.498	1.324	1.915	2.493	3.246
4	.525	1.329	1.894	2.438	3.139

Criteria met when |t.obs| < ADK Criteria(%-tile of interest)

Also: P-value should be > α (rule of thumb)

Test Set-up Experiences

- Test bed set up may have to be described in more detail.
- We’ ve worked with a single vendor.

Selecting the proper Operation System took us one week (make sure support of L2TPv3 is a main purpose of that software).

Connect the IPPM implementation to a switch and install a cable or internal U-turn on that switch. Maintain separate IEEE 802.1q logical VLAN connections when connecting the switch to the CPE which terminates the L2TPv3 tunnel.

The CPE requires at least a route-able IP address as LB0 interface, if the L2TPv3 tunnel spans the Internet.

The Ethernet Interface MUST be cross connected to the L2TPv3 tunnel in port mode.

- Terminate the L2TPv3 tunnel at the LB0 interface.
- Don’t forget to configure firewalls and other middle boxes

Don’ t forget to configure firewalls and other middle boxes

NetProbe 5.8.5

Runs on Solaris (and Linux, occasionally)

- Pre-dates *WAMP, functionally similar
- Software-based packet generator

including Loss, Delay, PDV, Reordering,

Duplication, burst loss, etc. in post-processing on stored packet records

Duplication, burst loss, etc. in post-processing on stored packet records

- See Section 3.5 of [RFC2679], 3rd bullet point and also Section 3.8.2 of [RFC2679].

- 1. configure a path with 1 sec one-way constant delay

- 2. measure (average) one-way delay with 2 or more implementations, using identical waiting time thresholds for loss set at 2 seconds

- 3. configure the path with 3 sec one-way delay (or change the delay while test is in progress, measurements in step 2)

- 4. repeat measurements

- 5. observe that the increase measured in step 4 caused all packets to be declared lost, and that all packets that arrive

Section 6.3: First-bit to Last-bit

See Section 3.7.2 of [RFC2679], and Section 10.2 of [RFC2330].

See Section 3.7.2 of [RFC2679], and Section 10.2 of [RFC2330].

- 1. configure a path with 1000 ms one-way constant delay, and ideally including a low-speed link (10-baseT, FD)
- 2. measure (average) one-way delay with 2 or more implementations, using identical options and equal size small packets (e.g., 44 octet IP payload)

identical options and equal size small packets (e.g.,

- 3. maintain the same path with 1000 ms one-way delay
- 4. measure (average) one-way delay with 2 or more implementations, using identical options and equal size large packets (e.g., 480 octet IP payload)
- 5. observe that the increase measured in steps 2 and 4 is equivalent to the

Other Examples

6.4 One-way Delay, RFC 2679

- This test is intended to evaluate measurements in sections 3 and 4 of [RFC2679].

Average delays before/after 2 second increase

4. Error Calibration, RFC 2679

- This is a simple check to determine if an implementation reports the error calibration as required in Section 4.8 of [RFC2679].